УДК 004.89:004.93

М.С. Клименко
Інститут проблем штучного інтелекту МОН і НАН України, Україна
пр. академіка Глушкова, 40, м. Київ, 03680

МЕТОД РОЗПІЗНАВАННЯ ЕМОЦІЙНОГО СТАНУ ДИКТОРА ЗА ФРАЗОВИМИ МОДЕЛЯМИ

M.S. Klymenko
Institute of artificial intelligence problems of MES and NAS of Ukraine, Ukraine
40, Academician Hlushkov av., Kyiv, 03680

METHOD OF EMOTIONAL CONDITION RECOGNIZING BY VOICE USING PHRASES MODELS

У статті наведено метод розпізнавання емоційного стану за голосом, у якому ознакові описи емоційних станів формуються за множиною акустичних, просодичних та екстралінгвістичних характеристик. Запропоновано групування характеристик за їх розташуванням у типових фрагментах інтонаційних конструкцій. Числові дослідження показали, що даний підхід дозволив підвищити ймовірність розпізнавання емоцій емоційних станів порівняно із результатами без використання згрупованих ознак.

Ключові слова: модель диктора, акустичні характеристики емоцій, модель сумішей Гауса.

In the article the method of emotional condition recognizing by voice is described, in which models of emotional conditions are formed by a set of acoustic, prosodic and extra-linguistic characteristics. The grouping of characteristics by their placement in typical fragments of intonational constructions is proposed. Numeric researches have shown that this approach made it possible to increase the likelihood of emotional recognition of emotional states compared to the results without the use of grouped characteristics.

Keywords: speaker model, acoustic characteristics of emotions, the Gaussian mixture model.

Вступ

Задача автоматизованого розпізнавання емоційних станів людини на сьогоднішній день є актуальною у багатьох сферах розпізнавання образів. Складнощі виникають через не стаціонарність та нестабільність проявів емоцій, які, зазвичай, зовні слабо виражені й швидко змінюються. І якщо у безпосередньому контакті з людиною зафіксувати ознаки емоцій можна за допомогою виміру тиску, частоти скорочення серця та електромагнітної активності мозку, то віддалене розпізнавання проводиться лише шляхом фіксації візуальних (рухи, міміка) та звукових образів. У світі є певні досягнення у розпізнаванні низки емоцій за допомогою візуальних образів на сталих зображеннях, натомість про існування аналогічних розробок на основі звукових образів досі невідомо.

Метою даної статті є розробка методу, здатного виконувати розпізнавання у звуковому сигналі наявності проявів емоції у людини за її мовленнєвою активністю. В Україні задача розпізнавання емоцій за акустичними ознаками не набула розповсюдження, а відповідні запатентовані методи розпізнавання не є доступними для дослідження й використання.

Постановка задачі

Виходячи з мети роботи, були поставлені наступні задачі:

1. Проаналізувати перелік акустичних ознак, за якими можлива параметризація емоцій людини, на основі яких можна розробити та описати метод розпізнавання емоцій.
2. Відібрати перелік емоцій, прояви яких будуть розпізнаватись, та виконати їх ознаковий опис.

3. Чисельно дослідити ефективність запропонованого методу розпізнавання емоцій.

Аналіз акустичних ознак проявів емоцій

Виходячи із задачі, параметри мають бути якомога менше залежними від індивідуальних особливостей голосового тракту дикторів. Таким чином, недоцільно використовувати наступні засоби, що характеризують особливості диктора: висоту тону, абсолютну силу голосу, а також перехідні спектральні процеси у міжфонемних відрізках мовного сигналу [1]. Для виокремлення особливостей прояву емоційних станів було обрано наступні параметри.

1. Нормовані значення енергетичного спектра. Для отримання наборів даних інтегральних ознак на кожному з фреймів обчислюється короткочасний енергетичний спектр за допомогою фільтрації гребінкою цифрових фільтрів. Для обчислення смуг пропускання в рамках характеристик була використана барк-шкала, пов'язана з критичними смугами слуху, а також шкала півтонів натурального ладу. Такий вибір зумовлений психоакустичними принципами сприйняття.

Після фільтрації гребінкою з M цифрових фільтрів (залежно від використаної шкали) мовний сигнал може бути представлений у вигляді двовимірного масиву значень короткочасних енергетичних спектрів (спектральних зрізів), отриманих на кожному вікні аналізу:

$$\{x(1,j),\ldots, x(i,j),\ldots, x(M,j)\},\quad j=1,2,\ldots,J,$$

де $x(i,j)$ – значення енергії сигналу на виході i-го смугового фільтра у j-му спектральному зрізі;

J – загальна кількість вікон на відрізку сигналу.

Нормування значень ознак виконується для зниження залежності значень від лінійних викривлень мовного сигналу при його звукозаписі. Введемо операцію нормування масиву $\{a(i)\}_{i=1}^{N}$ по $2k+1$ точках:

$$d(a(i),k)=\begin{cases} d(a(k+1),k),\, i=\overline{1,k} \\ \dfrac{a(i)}{\sum\limits_{j=i-k}^{i+k} a(j)},\, i=\overline{k+1,N-k} \\ d(a(N-k),k),\, i=\overline{N-k+1,N} \end{cases}.$$

Тоді нормовані значення енергетичного спектра обчислюються як:

$$X(i)=\frac{x(i)}{\sum\limits_{i=1}^{M} x(i)},$$

де $x(i)$ – середнє значення по рядку масиву:

$$x(i)=\frac{1}{J}\sum_{j=1}^{J} x(i,j).$$

2. Відносний час перебування сигналу у смугах енергетичного спектра. Значення кожної i-го ознаки обчислюється за формулою:

$$t(i)=\frac{\Delta J(i)}{J},$$

де $\Delta J(i)$ – кількість спектральних зрізів, за яких енергія в i-й смузі перевищує середнє значення.

3. Відносна потужність спектра мовлення в смугах. Обчислюється за формулою:

$$P_H(i) = d\left(P(i), \mathrm{k}\right), P(i) = \frac{m(i)}{\Delta J(i)}.$$

Наведені вище інтегральні ознаки дають змогу апроксимувати особливості стану фільтруючих функцій мовного тракту, динаміка змін яких буде свідчити про наявність психоемоційних або фізіологічних збудників.

4. Значення компонент гістограми розподілу частоти основного тону.

Частота основного тону (ЧОТ) – значення частоти коливань голосових зв'язок у діапазоні 80-350 Гц (F_0), характеризує особливості голосу особи, відносно яких можливий аналіз динамічних характеристик. Для визначення значення ЧОТ у мовному сигналі застосовується наступний алгоритм. З відрізка сигналу за допомогою порогів усуваються фрагменти, що відповідають низькоенергетичним елементам мовлення, і ті ділянки, що мають високу частоту перетину нульового рівня сигналу (частіше за все, приголосні звуки). Отриманий таким чином сигнал розбивається на вікна і на кожному з вікон визначається ЧОТ за допомогою методу кепстрального аналізу, який зводиться до пошуку піку в області можливих значень ЧОТ, координата піку дає оцінку періоду даної частоти.

Дані ознаки призначені для опису особливостей розподілу значень основного тону голосу людини в діапазоні 50-400 Гц. Наступні компоненти гістограми обрані для опису розподілу ЧОТ: значення середньої, максимальної та мінімальної частоти, асиметрія та ексцес щільності розподілу.

5. *Кепстральні коефіцієнти*. Для відокремлення сигналу збудження від сигналу мовного тракту вдаються до кепстрального аналізу. Мел-частотні кепстральні коефіцієнти враховують психоакустичні принципи сприйняття мови, оскільки використовують шкалу Мел, пов'язану з критичними смугами. Для шкали Мел межі смуг відповідають центральним частотам Барк шкали. У даній роботі використано 13 трикутних фільтрів, розташованих рівномірно по шкалі Мел від 0 до частоти Найквіста. Цього достатньо для охоплення смуги частот мовного сигналу із необхідною роздільною здатністю.

6. *Просодичні характеристики.* Основними компонентами просодичного аналізу є інтонація і наголос. Фізично інтонація і наголос реалізуються сукупністю акустичних засобів (просодичних характеристик мови), до числа яких відносяться:
- мелодика – рух частоти основного тону (F_0);
- ритміка – поточна зміна тривалості звуків і пауз;
- енергетика – поточна зміна сили (амплітуди) звуку.

Головним компонентом інтонації є мелодика, яка описується мелодійним контуром. Мелодійний контур – характерна для мови картина зміни основного тону, звільнена від сегментних і позиційних впливів. Необхідно враховувати сукупність характеристик мелодійного контура для створення повного уявлення про певну *інтонаційну конструкцію*. Основні параметри мелодійного контура, які необхідні для визначення типу інтонаційної конструкції:
- початкова та кінцева частоти (значення ЧОТ першого та останнього відліків фрагменту контура);
- максимальна та мінімальна частоти (максимальне та мінімальне значення частоти ОТ у межах контура);

- середня частота (усереднене значення частоти ОТ у межах контуру);
- час максимуму та мінімуму (позиції максимального та мінімального значень ЧОТ у відсотках від довжини усього фрагмента);
- час половини частоти (позиція значення середньої ЧОТ у відсотках від довжини усього фрагмента);
- швидкість зміни тону (середня швидкість зростання чи спаду тону на відрізку, Гц/мс).

За допомогою цього набору просодичних характеристик визначається тип інтонаційної конструкції речення. В основі методики визначення типу інтонаційної конструкції, що використовується в даній роботі, лежить модель інтонаційних портретів акцентних одиниць [2]. Основні ознаки інтонаційних конструкцій наведені у таблиці 1.

Таблиця 1. Типи інтонаційних конструкцій висловлювань

№	Тип інтонаційної конструкції висловлювання	Напрям тону
1.	Розповідне речення	Спадаючий на кінці
2.	Спеціальне питання, веління	Спадаючий або сильно спадаючий
3.	Загальне питання, незавершеність	Зростаючий
4.	Порівняльне питання	Спадаючий або спадаюче-зростаючий
5.	Вигук	1-й центр – зростаючий
6.	Оціночний вигук	2-й центр – зростаючий
7.	Експресивна оцінка	Зростаючий до середини, потім спадаючий

7. *Екстралінгвістичні події*. Наявність у фрагментах пауз мовленнєвого сигналу кашлю, зітхань, плачу, сміху або інших, з акустичних подій, притаманних певним емоціям, частіше за все не беруться до уваги. Ці події називаються екстралінгвістичними або позамовними, оскільки не є результатом мовного процесу. У даній роботі для автоматизації пошуку таких подій запропоновано створити окремі моделі екстралінгвістичних подій через їх особливість прояву, не схожу на мовну активність.

Використання як ознак характеристик різного роду накладає умови використання універсальних методів узагальнення. Для даної задачі використаємо метод сумішей Гауса, який не спирається на специфіку параметрів і може працювати з векторами ознак великого порядку [3].

Моделі, що створюються на основі сумішей Гауса, поділяють простір ознак на області, в яких сконцентровані значення векторів ознак. Класи в просторі ознак описуються у вигляді багатовимірного ймовірнісного розподілу. Основна ідея представити його у вигляді зваженої суми *M* нормальних розподілів:

$$p(\bar{x} \mid \lambda) = \sum_{i=1}^{M} w_i p_i(\bar{x}),$$

де $\bar{x}$ – *N*-вимірний вектор ознак;

w_i – вагові коефіцієнти компонентів моделі;

p_i – багатовимірні функції щільності розподілу складових моделі.

Таким чином, повністю модель описується векторами математичного очікування, коваріаційною матрицею і вагами сумішей для кожного компонента моделі.

Широко вживаним способом оцінки параметрів моделі є метод максимізації правдоподібності, функція якого має вигляд:

$$p(X \mid \lambda) = \prod_{t=1}^{T} p(x_t \mid \lambda),$$

де $X = \{\bar{x}_1, \ldots, \bar{x}_T\}$ – послідовність векторів ознак.

Припускаючи, що усі прояви емоцій однаково ймовірні, спрощене правило класифікації має вигляд:

$$res = \arg \max_{1 \le k \le S} p(X \mid \lambda_k),$$

де S – кількість емоційних проявів.

Ознаковий опис множини емоцій

На сьогодні не існує однієї загальноприйнятої класифікації людських емоцій. Загальновідомим є перелік 10 «фундаментальних» емоцій К. Ізарда [4]: інтерес, радість, здивування, страждання, гнів, відраза, зневага, страх, сором та провина. Оскільки у даній класифікації наявні близькі як за семантикою, так і за проявом емоції (інтерес та здивування, відраза та зневага, сором та провина), виключимо по одній емоції з цих пар, спростивши задачу розпізнавання у даній роботі. Розглянемо докладніше прояви обраних 7 емоцій.

Інтерес – почуття захопленості, цікавості, має яскравий стенічний характер. Перебування у стані носить стійкий інтенсивний характер. Прояв інтересу у голосі помітно по контуру речень, що зростають інтонаційно, а також по зменшенню пауз між словами до 12%. Окрім інтонаційних характеристик можна спостерігати статистично більше значення максимальної ЧОТ (до 6%) і відповідно незначну позитивну асиметрію щільності розподілу. На тестових зразках характерною є ознака відсутності екстралінгвістичних подій.

Радість – позитивна емоція, пов'язана з можливістю досить повно задовольнити актуальну потребу, ймовірність чого до цього була невелика або невизначена. Має стенічний характер. Радість характеризується збільшенням нормованого часу перебування сигналу у смугах середніх і високих частот енергетичного спектра мовлення (S_8-S_{17}). Фіксувалось збільшення нормованої енергії спектра. Швидкість вимови має зростання до 5%. Ймовірні екстралінгвістичні події: сміх, плач.

Страждання – негативний емоційний стан, пов'язаний з отриманням достовірної (чи яка здавалася б такою) інформації про неможливість задоволення найважливіших потреб, досягнення яких до цього оцінювалось більш ймовірним. Страждання має прояви, які характеризуються, перш за все, просодичними характеристиками: інтонаційний контур спадаючого типу або максимально рівний, негативна асиметрія щільності розподілу ЧОТ від -2% до -11%, ймовірні появи збільшеної тривалості пауз, а також зменшення швидкості вимови на 8%.

Гнів – виникає у відповідь на перешкоду в досягненні пристрасно бажаних цілей. Гнів має характер стенічної емоції. Прояви характеризуються збільшенням на 5-16% відносного часу перебування сигналу у смугах енергетичного спектра, близьких до ЧОТ (S_2-S_6). Відносна потужність спектра мовлення більша насамперед

у верхніх смугах частот (S_{10}-S_{17}). Швидкість вимови може зростати, а інтонаційні конструкції частіше за інших є типу вигук, оціночний вигук або експресивна оцінка. Ймовірні екстралінгвістичні події: крик, вигук, задишка.

Відраза – негативний емоційний стан, що викликається об’єктами, зіткнення з якими вступає в різке протиріччя з принципами та установками суб’єкта. Відраза має ознаки гніву, але відносний час перебування сигналу у смугах енергетичного спектра зростає незначно – до 4%. Характеристики ЧОТ відрізняються відсутністю збільшеного ексцесу. Щодо тривалості пауз та швидкості вимови – ознаки демонструють відхилення від стану спокою, але коливання можливі в обох напрямках.

Страх – негативний емоційний стан, що з’являється при отриманні суб’єктом інформації про можливу чи уявну небезпеку для себе або об’єктів, які мають високу ціну для суб’єкта. Проявляється відхиленням параметра нормованих значень енергетичного спектра у смугах із максимальним скупченням потужності мовлення (S_4-S_{12}), а також короткочасними сильними підвищеннями відносної потужності спектра мовлення у верхніх смугах від 6% до 22%. Серед значень ЧОТ наявна значна негативна асиметрія щільності розподілу, а з екстралінгвістичних подій є ймовірною поява крику, плачу та задишки.

Сором – негативний емоційний стан, що виражається в усвідомленні невідповідності власних помислів, вчинків не тільки із очікуваннями оточуючих, а й з власними уявленнями про належну поведінку в даній ситуації. Має астенічний характер та інтенсивність проявів від низької до середньої. Прояви сорому можуть бути параметризовані зменшенням на 2-6% відносного часу перебування сигналу у нижніх смугах енергетичного спектра (S_3-S_8). Із просодичних характеристик сорому можна виділити зменшення діапазону зміни ЧОТ, ексцесу щільності розподілу її значень до 7%. Тривалість пауз може зростати від 5% до 16%, а інтонаційні конструкції частіше за інших мають рівний характер або із зростанням до кінця.

Числове дослідження ефективності методу

Для проведення числового дослідження було виконано запис фрагментів прояву 7 обраних емоцій. Запис виконувався у моно режимі із частотою дискретизації 44100 Гц, глибиною квантування 16 біт, а для збереження використовувався формат WAV PCM без втрати якості. У записі брали участь 20 дикторів (12 чоловічої та 8 жіночої статі) віком від 22 до 38 років. Кожним диктором було виконано імітацію проявів 7 емоцій. Запис прояву однієї емоції тривав 16-35 с., що дозволило отримати по 1-2 фрагменти окремих проявів із завершеною інтонаційною конструкцією для можливості отримання усіх зазначених характеристик. Результати розпізнавання методом без урахування особливостей інтонаційного контуру наведено у таблиці 2.

Таблиця 2. Результати розпізнавання проявів емоцій без урахування типу ІК

Емоція, що розпізнається		Результат розпізнавання прояву емоції						
№	Назва	1	2	3	4	5	6	7
1	Інтерес	13	4			2	1	
2	Радість	3	15		1		1	
3	Страждання		2	12		3	2	1
4	Гнів		3		15		2	
5	Відраза			3	2	10	2	3
6	Страх		2		4		14	
7	Сором			4	2	2		12

Проаналізувавши табличні дані, стає зрозумілою певна згрупованість помилок розпізнавання серед підмножин емоцій, а саме інтересу та радості, страждання та відрази, страху та гніву. Це є свідченням відсутності чіткої параметризації даних емоцій наявними ознаками.

У таблиці 3 наведено результати низки розпізнавань запропонованим методом із групуванням ознак за фрагментами інтонаційного контуру.

Таблиця 3. Результати розпізнавання проявів емоцій за допомогою запропонованого методу

Емоція, що розпізнається		Результат розпізнавання прояву емоції						
№	Назва	1	2	3	4	5	6	7
1	Інтерес	16	3				1	
2	Радість	2	18					
3	Страждання			17		2	1	
4	Гнів			2	18			
5	Відраза	1		2		16		1
6	Страх			2	1		17	
7	Сором			3		1		16

Ймовірність правильного розпізнавання кожної емоції у даному випадку склала не менше 80%. Скоротилась не тільки ймовірнісна характеристика помилки, а й кількісна – множина помилкових варіантів емоцій. Досі простежується ефект згрупованості помилок результатів першого дослідження, що свідчить про необхідність вдосконалення розділової здатності ознак для врахування особливостей проявів даних емоцій.

Висновки

У статті запропоновано метод розпізнавання емоцій за голосом, який використовує особливості інтонаційного контуру для групування ознак. Даний підхід дозволив отримати більш деталізовану інформацію щодо особливостей стану голосу, які набувають стаціонарного характеру саме в межах інтонаційного фрагменту.

Для числових досліджень ефективності методу обрано 7 емоцій, які мають більшу відносно інших амплітуду зміни проявів у голосі. Для їх ознакового опису використано множину із 7 характеристик різного роду, які дозволяють створити ознаковий простір для з’ясування особливостей прояву емоцій. Числове дослідження якості автоматичного розпізнавання за усією множиною характеристик показало середню ймовірність розпізнавання емоцій – 84% серед фрагментів 20 дикторів. Складність для класифікації являють собою групи емоцій, що мають близькі параметри ознак. Вирішення цих проблем може стати подальшим розвитком даної роботи.

Література

1. Клименко Н.С. Исследование эффективности бустинга в задаче текстонезависимой идентификации диктора / Н.С. Клименко, И.Г. Герасимов // Штучний інтелект. – 2014. – №4 (66). – С. 191-201.
2. Лобанов Б.М. Компьтерный синтез и клонирование речи / Б.М. Лобанов, Л.И. Цирульник. – Минск: «Белорусская наука». – 2008. – 316 с.
3. Figueiredo M.A.T. Unsupervised Learning of Finite Mixture Models / M.A.T. Figueiredo, A.K. Jain // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2002. – №24 (3). – P. 381–396.
4. Психология эмоций / Изард К.Э. Перев. с англ. - СПб: Издательство "Питер", 1999. - 464 с.

Literatura

1. Klimenko N.S. Issledovanie effektivnosti bustinga v zadache tekstonezavisimoy identifikatsii diktora / N.S. Klimenko, I.G. Gerasimov // Shtuchniy Intelekt. – 2014. – №4 (66). – C. 191-201.
2. Lobanov B.M. Kompternyiy sintez i klonirovanie rechi / B.M. Lobanov, L.I. Tsirulnik. – Minsk: "Belorusskaya nauka". – 2008. – 316 s.
3. Figueiredo M.A.T. Unsupervised Learning of Finite Mixture Models / M.A.T. Figueiredo, A.K. Jain // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2002. – №24 (3). – P. 381–396.
4. Psihologiya emotsiy / Izard K.E. Perev. s angl. - SPb: Izdatelstvo "Piter", 1999. - 464 s.

RESUME

M.S. Klymenko

Method of emotional condition recognizing by voice using phrases models

The article describes the method of emotional condition recognizing by voice. To distinguish the features of emotional conditions, an analysis of characteristics suitable for a given task was conducted, from which a set of acoustic, prosodic and extra-linguistic characteristics were chosen. The development of the method was carried out by analyzing the values of these characteristics, which are calculated on the sound fragments of emotional conditions manifestations.

Analysis and numeric researches were performed on the following set of emotional conditions: interest, joy, suffering, anger, aversion, fear and shame. These emotional conditions were chosen from a wider list, except for relatives, both in terms of semantics and manifestation.

In the analysis of the distribution of the values of characteristics, an increase in the density of distribution in places with a certain type of fragment of the intonation contour was observed. This property was the basis of the method. Its essence is that the feature vectors of all characteristics are grouped by their belonging to fragments of the intonation contour, which is calculated first of all. Then, on the basis of each group of feature vectors, a model based on the Gauss mixture method is formed. The combination of such models describes a general model of emotional condition manifestation.

The numeric researches among the fragments of 20 speakers for the whole set of characteristics showed an average probability of emotion recognition of 84%, which exceeds the similar results of recognition without grouping of feature vectors.

Надійшла до редакції 10.08.2017